Basics of Molecular Biology

Martin Tompa
Department of Computer Science and Engineering
Department of Genome Sciences
University of Washington
Seattle, WA 98195-2350
U.S.A.

July 6, 2003
Updated December 18, 2009

We begin with a review of the basic molecules responsible for the functioning of all organisms' cells. Much of the material here comes from the introductory textbooks by Drlica [4], Lewin [7], and Watson *et al.* [10]. Good short primers have been written by Hunter [6] and Brāzma *et al.* [2].

What sorts of molecules perform the required functions of the cells of organisms? Cells have a basic tension in the roles they need those molecules to fulfill:

1. The molecules must perform the wide variety of chemical reactions necessary for life. To perform these reactions, cells need diverse three-dimensional structures of interacting molecules.

2. The molecules must pass on the instructions for creating their constituent components to their descendents. For this purpose, a simple one-dimensional information storage medium is the most effective.

We will see that *proteins* provide the three-dimensional diversity required by the first role, and *DNA* provides the one-dimensional information storage required by the second. Another cellular molecule, *RNA*, is an intermediary between DNA and proteins, and plays some of each of these two roles.

1 Proteins

Proteins have a variety of roles that they must fulfill:

1. They are the enzymes that rearrange chemical bonds.

2. They carry signals to and from the outside of the cell, and within the cell.

3. They transport small molecules.

4. They form many of the cellular structures.

5. They regulate cell processes, turning them on and off and controlling their rates.

This variety of roles is accomplished by the variety of proteins, which collectively can assume a variety of three-dimensional shapes.

A protein's three-dimensional shape, in turn, is determined by the particular one-dimensional composition of the protein. Each protein is a linear sequence made of smaller constituent molecules called *amino acids*. The constituent amino acids are joined by a "backbone" composed of a regularly repeating sequence of bonds. (See [7, Figure 1.4].) There is an asymmetric orientation to this backbone imposed by its chemical structure: one end is called the *N-terminus* and the other end the *C-terminus*. This orientation imposes directionality on the amino acid sequence.

There are 20 different types of amino acids. The three-dimensional shape the protein assumes is determined by the specific linear sequence of amino acids from N-terminus to C-terminus. Different sequences of amino acids *fold* into different three-dimensional shapes. (See, for example, [1, Figure 1.1].)

Protein size is usually measured in terms of the number of amino acids that comprise it. Proteins can range from fewer than 20 to more than 5000 amino acids in length, although an average protein is about 350 amino acids in length.

Each protein that an organism can produce is encoded in a piece of the DNA called a "gene" (see Section 6). To give an idea of the variety of proteins one organism can produce, the single-celled bacterium *E. coli* has about 4300 different genes. Humans are believed to have about 25,000 different genes (the exact number as yet unresolved), so a human has only about 6 times as many genes as *E. coli*. The number of proteins that can be produced by humans greatly exceeds the number of genes, however, because a substantial fraction of the human genes can each produce many different proteins through a process called "alternative splicing".

1.1 Classification of the Amino Acids

Each of the 20 amino acids consists of two parts:

1. a part that is identical among all 20 amino acids; this part is used to link one amino acid to another to form the backbone of the protein.
2. a unique *side chain* (or "R group") that determines the distinctive physical and chemical properties of the amino acid.

Although each of the 20 different amino acids has unique properties, they can be classified into four categories based upon their major chemical properties. Below are the names of the amino acids, their 3 letter abbreviations, and their standard one letter symbols.

1. Positively charged (and therefore basic) amino acids (3).

Arginine	Arg	R
Histidine	His	H
Lysine	Lys	K

2. Negatively charged (and therefore acidic) amino acids (2).

Aspartic acid	Asp	D
Glutamic acid	Glu	E

3. Polar amino acids (7). Though uncharged overall, these amino acids have an uneven charge distribution. Because of this uneven charge distribution, these amino acids can form hydrogen bonds with water. As a consequence, polar amino acids are often found on the outer surface of folded proteins, in contact with the watery environment of the cell, in which case they are called *hydrophilic*.

Asparagine	Asn	N
Cysteine	Cys	C
Glutamine	Gln	Q
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y

4. Nonpolar amino acids (8). These amino acids are uncharged and have a uniform charge distribution. Because of this, they do not form hydrogen bonds with water, and tend to be found on the inside surface of folded proteins, in which case they are called *hydrophobic*.

Alanine	Ala	A
Isoleucine	Ile	I
Glycine	Gly	G
Leucine	Leu	L
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Valine	Val	V

This classification of the physio-chemical properties of the amino acids is overly simplistic. A more accurate depiction of their properties is given in the Venn diagram of Livingstone and Barton [8] at `http://www.russell.embl-heidelberg.de/aas/aas.html`.

Although each amino acid is different and has unique properties, certain pairs have more similar properties than others. The two nonpolar amino acids leucine and isoleucine, for example, are far more similar to each other in their chemical and physical properties than either is to the charged glutamic acid. In algorithms for comparing proteins, the question of amino acid similarity will be important.

2 DNA

DNA contains the instructions needed by the cell to carry out its functions. DNA consists of two long interwoven strands that form the famous "double helix". (See [4, Figure 3-3].) Each strand is built from a small set of constituent molecules called *nucleotides*.

2.1 Structure of a Nucleotide

A nucleotide consist of three parts [4, Figure 3-2]. The first two parts are used to form the ribbon-like backbone of the DNA strand, and are identical in all nucleotides. These two parts are (1) a *phosphate group* and (2) a sugar called *deoxyribose* (from which DNA, DeoxyriboNucleic Acid, gets its name). The third part of the nucleotide is the *base*. There are four different bases, which define the four different nucleotides: thymine (T), cytosine (C), adenine (A), and guanine (G).

Note in [4, Figure 3-2] that the five carbon atoms of the sugar molecule are numbered $C_{1'}, C_{2'}, C_{3'}, C_{4'}, C_{5'}$. The base is attached to the $1'$ carbon. The two neighboring phosphate groups are attached to the $5'$ and $3'$ carbons. As is the case in the protein backbone (Section 1), the asymmetry of the sugar molecule imposes an orientation on the backbone, one end of which is called the *$5'$ end* and the other the *$3'$ end*. (See [4, Figure 3-4(a)].)

2.2 Base Pair Complementarity

Why is DNA double-stranded? This is due to *base pair complementarity*. If specific bases of one strand are aligned with specific bases on the other strand, the aligned bases can *hybridize* via hydrogen bonds, weak attractive forces between hydrogen and either nitrogen or oxygen. The specific complementary pairs are

- A with T
- G with C

Two hydrogen bonds form between A and T, whereas three form between C and G. (See [4, Figure 3-5].) This makes C-G bonds stronger than A-T bonds.

If two DNA strands consist of complementary bases, under "normal" cellular conditions they will hybridize and form a stable double helix. However, the two strands will only hybridize if they are in "antiparallel configuration". This means that the sequence of one strand, when read from the $5'$ end to the $3'$ end, must be complementary, base for base, to the sequence of the other strand read from $3'$ to $5'$. (See [4, Figure 3-4(b) and 3-3].)

2.3 Size of DNA molecules

An *E. coli* bacterium contains one circular, double-stranded molecule of DNA consisting of approximately 5 million nucleotides. Often the length of double-stranded DNA is expressed in the units of basepairs (bp), kilobasepairs (Kb), or megabasepairs (Mb), so that this size could be expressed equivalently as 5×10^6 bp, 5000 Kb, or 5 Mb.

Each human cell contains 23 pairs of *chromosomes*, each of which is a long, double-stranded DNA molecule. Collectively, the 23 distinct chromosomes in one human cell consist of approximately 3×10^9 bp of DNA. Note that a human has about 1000 times more DNA than *E. coli* does, yet only about 10 times as many genes. (See Section 1.) The reason for this will be explained shortly.

3 RNA

Chemically, RNA is very similar to DNA. There are two main differences:

1. RNA uses the sugar *ribose* instead of deoxyribose in its backbone (from which RNA, RiboNucleic Acid, gets its name).
2. RNA uses the base uracil (U) instead of thymine (T). U is chemically similar to T, and in particular is also complementary to A.

RNA has two properties important for our purposes. First, it tends to be single-stranded in its "normal" cellular state. Second, because RNA (like DNA) has base-pairing capability, it often forms intramolecular hydrogen bonds, partially hybridizing to itself. Because of this, RNA, like proteins, can fold into complex three-dimensional shapes. (For an example, see `http://www.ibc.wustl.edu/~zuker/rna/hammerhead.html`.)

RNA has some of the properties of both DNA and proteins. It has the same information storage capability as DNA due to its sequence of nucleotides. But its ability to form three-dimensional structures allows it to have enzymatic properties like those of proteins. Because of this dual functionality of RNA, it has been conjectured that life may have originated from RNA alone, DNA and proteins having evolved later.

4 Residues

The term *residue* refers to either a single base constituent from a nucleotide sequence, or a single amino acid constituent from a protein. This is a useful term when one wants to speak collectively about these two types of biological sequences.

5 DNA Replication

What is the purpose of double-strandedness in DNA? One answer is that this redundancy of information is key to how the one-dimensional instructions of the cell are passed on to its descendant cells. During the cell cycle, the DNA double strand is split into its two separate strands. As it is split, each individual strand is used as a template to synthesize its complementary strand, to which it hybridizes. (See [4, Figure 5-2 and 5-1].) The result is two exact copies of the original double-stranded DNA.

In more detail, an enzymatic protein called *DNA polymerase* splits the DNA double strand and synthesizes the complementary strand of DNA. It synthesizes this complementary strand by adding *free nucleotides* available in the cell onto the $3'$ end of the new strand being synthesized [4, Figure 5-3]. The DNA polymerase will only add a nucleotide if it is complementary to the opposing base on the template strand. Because the DNA polymerase can only add new nucleotides to the $3'$ end of a DNA strand (i.e., it can only synthesize DNA in the $5'$ to $3'$ direction), the actual mechanism of copying both strands is somewhat more complicated. One strand can be synthesized continuously in the $5'$ to $3'$ direction. The other strand must be synthesized in short $5'$-to-$3'$ fragments. Another enzymatic protein, *DNA ligase*, glues these synthesized fragments together into a single long DNA molecule. (See [4, Figure 5-4].)

6 Synthesis of RNA and Proteins

The one-dimensional storage of DNA contains the information needed by the cell to produce all its RNA and proteins. In this section, we describe how the information is encoded, and how these molecules are synthesized.

Proteins are synthesized in a two-step process. First, an RNA "copy" of a portion of the DNA is synthesized in a process called *transcription*, described in Section 6.1. Second, this RNA sequence is read and interpreted to synthesize a protein in a process called *translation*, described in Section 6.2. Together, these two steps are called *gene expression*.

A *gene* is a sequence of DNA that encodes a protein or an RNA molecule. Gene structure and the exact expression process are somewhat dependent on the organism in question. The *prokaryotes*, which consist of the *bacteria* and the *archaea*, are single-celled organisms lacking nuclei. Because prokaryotes have the simplest gene structure and gene expression process, we will start with them. The *eukaryotes*, which include plants and animals, have a somewhat more complex gene structure that we will discuss after.

6.1 Transcription in Prokaryotes

How do prokaryotes synthesize RNA from DNA? This process, called transcription, is similar to the way DNA is replicated (Section 5). An enzyme called *RNA polymerase*, copies one strand of the DNA gene into a *messenger RNA* (*mRNA*), sometimes called the *transcript*. The RNA polymerase temporarily splits the double-stranded DNA, and uses one strand as a template to build the complementary strand of RNA. (See [4, Figure 4-1].) It incorporates U opposite A, A opposite T, G opposite C, and C opposite G. The RNA polymerase begins this transcription at a short DNA pattern it recognizes called the *transcription start site*. When the polymerase reaches another DNA sequence called the *transcription stop site*, signalling the end of the gene, it drops off.

6.2 Translation

How is protein synthesized from mRNA? This process, called translation, is not as simple as transcription, because it proceeds from a 4 letter alphabet to the 20 letter alphabet of proteins. Because there is not a one-to-one correspondence between the two alphabets, amino acids are encoded by consecutive sequences of 3 nucleotides, called *codons*. (Taking 2 nucleotides at a time would give only $4^2 = 16$ possible permutations, whereas taking 3 nucleotides yields $4^3 = 64$ possible permutations, more than sufficient to encode the 20 different amino acids.) The decoding table is given in Table 1, and is called the *genetic code*. It is rather amazing that this same code is used almost universally by all organisms.

	U	C	A	G	
U	UUU Phe [F]	UCU Ser [S]	UAU Tyr [Y]	UGU Cys [C]	U
	UUC Phe [F]	UCC Ser [S]	UAC Tyr [Y]	UGC Cys [C]	C
	UUA Leu [L]	UCA Ser [S]	UAA *STOP*	UGA *STOP*	A
	UUG Leu [L]	UCG Ser [S]	UAG *STOP*	UGG Trp [W]	G
C	CUU Leu [L]	CCU Pro [P]	CAU His [H]	CGU Arg [R]	U
	CUC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CUA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CUG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	AUU Ile [I]	ACU Thr [T]	AAU Asn [N]	AGU Ser [S]	U
	AUC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	AUA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	AUG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GUU Val [V]	GCU Ala [A]	GAU Asp [D]	GGU Gly [G]	U
	GUC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GUA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GUG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Table 1: The Genetic Code

There is a necessary redundancy in the code, since there are 64 possible codons and only 20 amino acids. Thus each amino acid (with the exceptions of Met and Trp) is encoded by *synonymous codons*, which are interchangeable in the sense of producing the same amino acid. Only 61 of the 64 codons are used to encode amino acids. The remaining 3, called *STOP codons*, signify the end of the protein.

Ribosomes are the molecular structures that read mRNA and produce the encoded protein according to the genetic code. Ribosomes are large complexes consisting of both proteins and a type of RNA called *ribosomal RNA* (*rRNA*).

The process by which ribosomes translate mRNA into protein makes use of yet a third type of RNA called *transfer RNA* (*tRNA*). There are 61 different transfer RNAs, one for each nontermination codon. Each tRNA folds (see Section 3) to form a cloverleaf-shaped structure. This structure produces a pocket that complexes uniquely with the amino acid encoded by the tRNA's associated codon, according to Table 1. The unique fit is accomplished analogously to a key and lock mechanism. Elsewhere on the tRNA is the *anticodon*, three consecutive bases that are complementary and antiparallel to the associated codon, and exposed for use by the ribosome. The ribosome brings together each codon of the mRNA with its corresponding anticodon on some tRNA, and hence its encoded amino acid. (See [4, Figure 4-4].)

In prokaryotes, which have no cell nucleus, translation begins while transcription is still in progress, the $5'$ end of the transcript being translated before the RNA polymerase has transcribed the $3'$ end. (See Drlica [4, Figure 4-4].) In eukaryotes, the DNA is inside the nucleus, whereas the ribosomes are in the *cytoplasm* outside the nucleus. Hence, transcription takes place in the nucleus, the completed transcript is exported from the nucleus, and translation then takes place in the cytoplasm.

The ribosome forms a complex near the $5'$ end of the mRNA, binding around the *start codon*, also called the *translation start site*. The start codon is most often $5'$-AUG-$3'$, and the corresponding anticodon is $5'$-CAU-$3'$. (Less often, the start codon is $5'$-GUG-$3'$or $5'$-UUG-$3'$.) The ribosome now brings together this start codon on the mRNA and its exposed anticodon on the corresponding tRNA, which hybridize to each other. (See [4, Figure 4-4].) The tRNA brings with it the encoded amino acid; in the case of the usual start codon $5'$-AUG-$3'$, this is methionine.

Having incorporated the first amino acid of the synthesized protein, the ribosome shifts the mRNA three bases to the next codon. A second tRNA complexed with its specific amino acid hybridizes to the second codon via its anticodon, and the ribosome bonds this second amino acid to the first. At this point the ribosome releases the first tRNA, moves on to the third codon, and repeats. (See [4, Figure 4-5].) This process continues until the ribosome detects one of the STOP codons, at which point it releases the mRNA and the completed protein.

7 Prokaryotic Gene Structure

Recall from Section 6 that a gene is a relatively short sequence of DNA that encodes a protein or RNA molecule. In this section we restrict our attention to protein-coding genes in prokaryotes.

The portion of the gene containing the codons that ultimately will be translated into the protein is called the *coding region*, or *open reading frame*. The transcription start site (see Section 6.1) is somewhat *upstream* from the start codon, where "upstream" means "in the $5'$ direction". Similarly, the transcription stop site is somewhat *downstream* from the stop codon, where "downstream" means "in the $3'$ direction". That is, the mRNA transcript contains sequence at both its ends that has been transcribed, but will not be translated. The sequence between the transcription start site and the start codon is called the *$5'$ untranslated region*. The sequence between the stop codon and the transcription stop site is called the *$3'$ untranslated region*.

Upstream from the transcription start site is a relatively short sequence of DNA called the *regulatory region* or *promoter region*. It contains *regulatory elements*, which are specific DNA sites where certain regulatory proteins bind and regulate expression of the gene. These proteins are called *transcription factors*, since they regulate the transcription process. A common way in which transcription factors regulate expression is to bind to the DNA at a promoter and from there affect the ability (either positively or negatively) of RNA polymerase to perform its task of transcription. (There is also the analogous possibility of *translational regulation*, in which regulatory factors bind to the mRNA and affect the ability of the ribosome to perform its task of translation.)

8 Prokaryotic Genome Organization

The *genome* of an organism is the entire complement of DNA in any of its cells. In prokaryotes, the genome typically consists of a single chromosome of double-stranded DNA, and it is often circularized (its $5'$ and $3'$ ends attached) as opposed to being linear. A typical prokaryotic genome size would be in the millions of base pairs.

Typically 85% of the prokaryotic genome consists of protein-coding regions. For instance, the *E. coli* genome has size about 5 Mb and approximately 4300 coding regions, each of average length around 1000 bp. The genes are relatively densely and uniformly distributed throughout the genome.

9 Eukaryotic Gene Structure

An important difference between prokaryotic and eukaryotic genes is that the latter may contain "introns". In more detail, the transcribed sequence of a general eukaryotic gene is an alternation between DNA sequences called *exons* and *introns*, where the introns are sequences that ultimately will be spliced out of the mRNA before it leaves the nucleus. Transcription in the nucleus produces an RNA molecule called *pre-mRNA*, produced as described in Section 6.1, that contains both the exons and introns. The introns are spliced out of the pre-mRNA by structures called *spliceosomes* to produce the *mature mRNA* that will be transported out of the nucleus for translation. A eukaryotic gene may contain numerous introns, and each intron may be many kilobases in size. One fact that is relevant to our computational gene prediction is that the presence of introns makes it much more difficult to identify the locations of genes computationally, given the genome sequence.

Another important difference between prokaryotic and higher eukaryotic genes is that, in the latter, there can be multiple regulatory regions that can be quite far from the coding region, can be either upstream or downstream from it, and can even be in the introns.

10 Eukaryotic Genome Organization

Unlike prokaryotic genomes, many eukaryotic genomes consist of multiple linear chromosomes as opposed to single circular chromosomes. Depending on how simple the eukaryote is, very little of the genome may be coding sequence. In humans, approximately 1.5% of the genome is believed to be protein-coding sequence, and the genes are distributed quite nonuniformly over the genome.

11 Goals and Status of Genome Projects

Molecular biology has the following two broad goals:

1. Identify all key molecules of a given organism, particularly the proteins, since they are responsible for the chemical reactions of the cells.
2. Identify all key interactions among molecules.

Traditionally, molecular biologists have tackled these two goals simultaneously in selected small systems within selected model organisms. The genome projects today differ by focusing primarily on the first goal, but for *all* the systems of a given model organism. They do this by *sequencing* the genome, which means determining the entire DNA sequence of the organism. They then perform a computational analysis on the genome sequence to identify (most of) the genes. Having done this, (many of) the proteins of the organism will have been identified.

With recent advances in sequencing technology, the genome projects have progressed very rapidly over the past five years. The first free-living organism to be completely sequenced was the bacterium *H. influenzae* in 1995 [5], with a genome of size 1.8 Mb. At the time of this writing, over 950 bacterial, 68 archaeal, and approximately 45 vertebrate genomes have been sequenced, plus numerous plants, insects, fungi, etc. (See `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome` for a current listing of sequenced genomes.)

The human genome was sequenced around 2001 [3, 9]. Although every human is a unique individual, the genome sequences of any two humans are about 99.9% identical, so that it makes some sense to talk about sequencing *the* human genome, which will really be an amalgamation of a small collection of individuals. Once that is done, one of the interesting challenges is to identify the common *polymorphisms*, which are genomic variations that occur in a nonnegligible fraction of the population.

12 Sequence Analysis

Once a genome is completely sequenced, what sorts of analyses are performed on it? Some of the goals of *sequence analysis* are the following:

1. Identify the genes.

2. Determine the function of each gene. One way to hypothesize the function is to find another gene (possibly from another organism) whose function is known and to which the new gene has high sequence similarity. This assumes that sequence similarity implies functional similarity, which may or may not be true.

3. Identify the proteins involved in the regulation of gene expression.

4. Identify sequence repeats.

5. Identify other functional regions, for example *origins of replication* (sites at which DNA polymerase binds and begins replication; see Section 5), *pseudogenes* (sequences that look like genes but are not expressed), sequences responsible for the compact folding of DNA, and sequences responsible for nuclear anchoring of the DNA.

Many of these tasks are computational in nature. Given the incredible rate at which sequence data is being produced, the integration of computer science, mathematics, and biology will be integral to analyzing those sequences.

References

[1] C. Branden and J. Tooze. *An Introduction to Protein Structure*. Garland, 1998.

[2] Alvis Brāzma, Helen Parkinson, Thomas Schlitt, and Mohammadreza Shojatalab. A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays, 2001. `http://www.ebi.ac.uk/microarray/biology_intro.html`.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, February 2001.

[4] Karl Drlica. *Understanding DNA and Gene Cloning*. John Wiley & Sons, second edition, 1992.

[5] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae rd. Science*, 269:496–512, July 1995.

[6] Lawrence Hunter. Molecular biology for computer scientists. In Lawrence Hunter, editor, *Artificial Intelligence and Molecular Biology*, chapter 1, pages 1–46. AAAI Press, 1993. `http://www.aaai.org//Library/Books/Hunter/01-Hunter.pdf`.

[7] Benjamin Lewin. *Genes VI*. Oxford University Press, 1997.

[8] C.D. Livingstone and G.J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer Applications in the Biosciences*, 9(6):745–756, December 1993.

[9] J. Craig Venter, Mark D. Adams, Eugene W. Myers, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, February 2001.

[10] James D. Watson, Michael Gilman, Jan Witkowski, and Mark Zoller. *Recombinant DNA*. Scientific American Books (Distributed by W. H. Freeman), second edition, 1992.